International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 5
May 2018

# Telugu Script Recognition Approach Using Kernel Features

**Kesana Mohana Lakshmi**[1]

[1]CMR Technical Campus, Kandlakoya,Medchal, Andhrapradesh, India

**Tummala Ranga Babu**[2]

[2]R.V.R & J.C College of Engineering, Guntur, Andhrapradesh, India

*ABSTRACT— Due to their many applications the optical character recognition (OCR) systems have been developed even for scripts like Telugu. Due to the huge number of symbols utilization, identifying the Telugu words are very much complicated. Pre-computed symbol features have been stored by these types of systems to be recognized or to retrieve in a database. Hence, searching of Telugu script from the database is a challenging task due to the complication in finding the features of the Telugu word images or scripts. Here, we had implemented novel Telugu script recognition based on the extraction of features for the TELUGU text, which uses KERNEL to extract text features and uses the AKD tree algorithm for matching purpose.*

***KEYWORDS: OCR, preprocessing, Kernel features, AKD***

## 1. INTRODUCTION

The process of automatic reading of documents is composed of a sequence of stages like image acquisition, pre-processing, object extraction, normalization or windowing, feature extraction, classification and post-processing. The image acquisition is the process of imbibing a document as an input to the character recognition system. Pre-processing stage subjects to removal of noises from the image that occurs due to variety of external factors like improper image scan settings, quality or resolution of image, quality of image capturing device and lack of illumination etc. The resulting images require additional pre-processing techniques like elimination of page layout [1] and graphical components in the image [2], skew detection and correction [3] etc followed by transforming the document to a suitable form for further processing. The object extraction generally called as segmentation; the process of identifying boundaries for region of interest (ROI). The offline OCRs imbibe a scanned document input and converts it into machine editable document format necessarily into Unicode of corresponding character images. The input documents are pre-composed with text of either printed or handwritten script pertaining to a particular language. The documents used in character recognition systems are classified as variety of types [4]. The standard classifications are printed and handwritten documents.

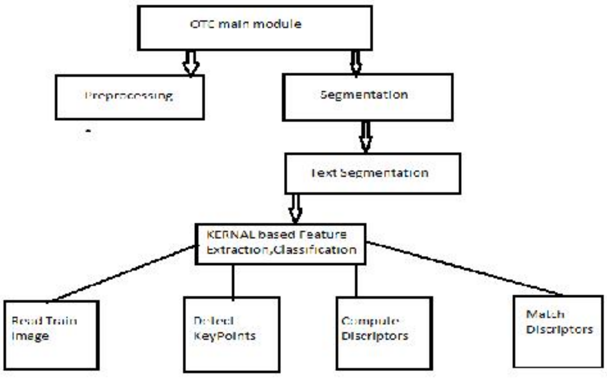


**Fig:1Modules and the flow of control in the OCR system**

The printed documents are composed with a particular font style and size of the language. The handwritten documents are written by the author in a particular script, handwriting style of the individual, with the freedom of writing. The freedom and flexibility taken by the author while writing the script adds more complexities in recognition of the characters, since each individual have their own handwriting style and in addition the mistakes while writing appends conflicts in the process of recognition. The combination of printed and handwritten document had resulted into another type of documents called as application/pre-printed documents or forms [5], consisting of both printed and handwritten scripts.

. [6] presented that to identify a word wise script; Gradient Local Auto-Correlation (GLAC) feature is very robust and effective and they found that for identifying or recognizing the Telugu script, the gradient feature is more suitable and effective that the traditional texture features. Author [7] studied and analyzed three various features named as Gabor, Zernike moments and gradient with 400 dimensions for word-wise script identification and classification has been done using Support Vector Machine (SVM). These authors have explained that the necessary pre-processing methods are required to overcome the problems with the input or source. A scheme named as template matching has been utilized for word recognition in [8] with the feature identification as Gradient Angular Features (GAF) tested on 760 words from six different scripts. Based on all the research papers discussed above and in the literature exposed that the unique modeling of structure of script is a challenging task for recognizing or identifying the word-script. Danish et. al. [9] employed the template matching technique for recognition of handwritten, machine printed and type written English characters and numerals obtained an accuracy of about 94.50% for standard typewritten fonts, 88% for unknown type written fonts, 98% for numerals and 75% in case of unknown type written fonts. Nikhil et. al. [10] applied the template for multi font styles and multi font sizes of English script and attained an accuracy of around 90%. Mo Wenying et. al. [11] applied the template matching algorithm by customizing with respect to weighted matching degree. This algorithm provides a higher matching rate and overcome the fallacious recognition produced by traditional calculation method with accuracy of around 100%. Jatin et. al. [12] employed the template matching technique for type written English characters and classified using neural network classifiers. Soumendu et. al. [13] had proposed an algorithm for Japanese character recognition using the centre of gravity features and Euclidean distance features and character with minimum Euclidean distance is the feature employed for character recognition.

This paper is organized as follows, section II explains about basics of KERNEL and then about AKD method. Section III explains about proposed methodology and IV gives the simulation results of the proposed methodology and followed by conclusion in section V

## II. KERNEL FEATURE AND AKD TREE

### A. KERNEL

Kernel forms maximum information in the image. Intuitively a kernel is a function that computes how similar two vectors are. The similar the vectors are to each other the higher the estimate of the kernel for those two vectors. We employ *kernel methods* to deal with the neural network's inefficiency and also to keep its advantages. Using a *kernel function*, the data is projected into a higher dimensional *feature space* that can possibly make the data linearly separable. Thus, the large number of training samples is made smaller since we utilize training algorithms like those of perception to learn the pixels *M and N*in this feature space. More importantly, the greatest benefit of using the kernel is many different kernel functions can be utilized within the same algorithm, and many different algorithms can be used with the same kernel function. We can easily change one from another based on the requirements.

### B. AKD TREE

The AKD distance between the reservation point and every maximum data point is calculated, and the closest neighbor pixel information is updated. The filter is most generally describe by substituting the values$V_i$ in a set of size n with a linear arrangement of all other values $V_j$ :

$$v_i = \sum_{j=1}^{n} \mathrm{W_{ii}}.\mathrm{V_j} \qquad \ldots\ldots \qquad (1)$$

We assume that values are represented by standardized coordinates, and the homogeneous neighbor information is filtered along with others.

$$v_i = \sum_{j=1}^{n} \mathrm{f}(|\mathrm{p_i}.\,\mathrm{p_j}|)\,.\,\mathrm{V_j} \;\ldots.. \qquad (2)$$

As it can be seen, kernel feature estimation works with a lower resolution image. It permits accurate matching of collections of functions in a high-dimensional appearance space but rejects all spatial invariance. This paper advocates an “AKD” method: execute features matching inside the two-dimensional image space and use conventional clustering techniques in feature space. Specifically, we quantize all feature matrix vectors into M discrete pixels and make the simplifying limitation that only functions of the matching type may be matched to each other. Each channel m gives us units of two-dimensional vectors, $X_m$ coefficients and $Y_m$ coefficients representing the coordinates of capabilities of type m discovered within the respective images.

$$(X \quad + Y \quad )^n = \sum_{k=0}^{n} \binom{n}{k} X^k a^{n-k} \quad \ldots\ldots\ldots\mathbf{(3)}$$

This segmentation, in brief, describes the 2 variety of functions used in the experiments. First, we have so known as “kernel features,” which might be orientated edge pixels, i.e points whose gradient (magnitude and direction) value in a given path exceeds a negligible threshold. We extract area factors at two scales and 8 orientations, for a total no. of channels M = 16.

$$f(x) = a_0 + \sum_{n=8}^{1} \left( a_n \,\mathrm{X}\frac{n}{L} + b_n \,\mathrm{Y}\frac{n}{L} \right) \mathbf{\text{-}(4)}$$

We designed these functions to gather an illustration much like the “AKD” or to word wide kernel features of the image. For higher discriminative strength, we also utilize upper dimensional, which can be SIFT descriptors is 4× 4-pixel patches computed over a grid with a spacing of 2 pixels. Our decision to apply which can be KERNEL descriptors An of 16× 16 pixel patches computed over a network with a spacing of 8 pixels. work higher for prospect classification.

.

### C. Algorithm

Step 1: Load the database script images

**Step 2:** Select and read a query image ‘Q’ from current directory

**Step 3:** Resize the ‘Q’

**Step 4:** Find the KERNEL Features of ‘Q’

**Step 5:** Now, read all the script images from the data base and resize the images with the size of ‘Q’

**Step 6:** Find the KERNEL Features of the database script images

**Step 7:** Now, calculate the similarity distance between ‘Q’ and data base script images

**Step 8:** Display the matched script image as a recognized script from the Database

## III. METHODOLOGY

This paper is focusing on four different operations, such as preprocessing, segmentation, feature extraction, and classification. This stage quickly depicts the pixel extraction method for word from the document image. In the first place, the word image is part into 16 x 16 patches. The higher dimensional KERNEL features[2] of 16 x 16-pixel patches are processed over a fix. That is, after structure separation of an image, 16 x 16 pixels around each matrix focus are careful to figure the kernel descriptors. Next, AKD-approach grouping strategy is executed at the patches (i.e. Filter descriptors of 64 estimations) from the preparation set for the area of the coefficients. The normal features length for our investigations is l024. Existing methods are SPM plan is employed to produce the element vector, that is then bolstered to the SVM classifier.

$$\mathrm{X}\,a \pm \mathrm{Y}\,b = 2\,\mathrm{Lxx}\frac{1}{2}(a \pm b)\,\mathrm{Lxy}\frac{1}{2}(a \mp b) -\mathbf{(5)}$$

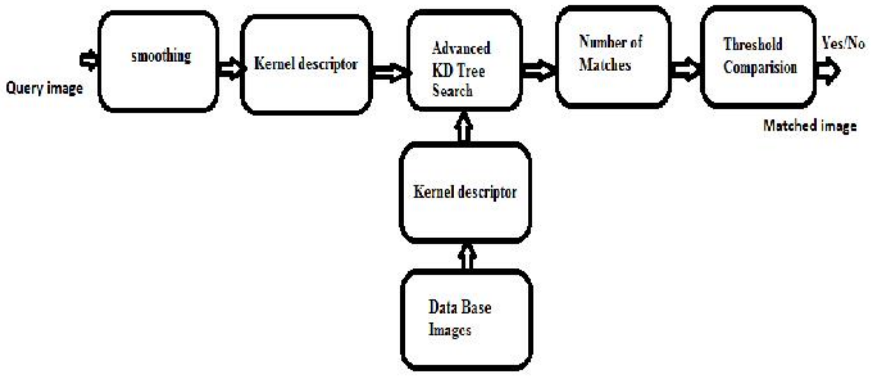


**Fig:2 Block Diagram Of Kernel Features**

**a) Smoothing**

Smoothing process filters all the unwanted data such as noise, blur, illumination etc, and get back the image in the original neutral form.

**c) Number of matches**

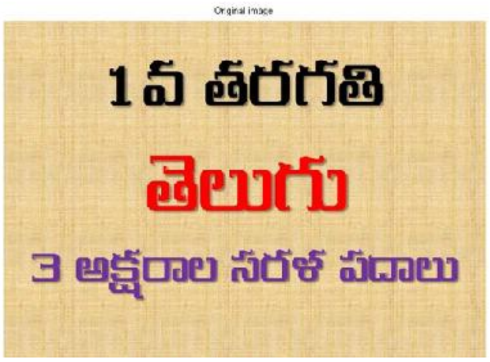
This block is concerned about keeping the record of the total amount of nodes getting matched between database image and trial image.

**d) Threshold Compression**

The value of each node of the trial data will be compared with the database images respectively. On this basis, the similarities between the database and test image are obtained which is much need for the reorganization of the data.

## IV. RESULTS

The proposed method, AKD with kernel features estimation has been implemented using MATLAB and special text and words also simulated. This simulation on different types of TELUGU scanned document images in the database and threshold for the matched lines is 0.9. The minimum value indicates the robustness of KERNEL. We found the good matches in accordance with similar features for query and database image and for different images, few matches are found.


Original image

1వ తరగతి

తెలుగు

3 అక్షరాల సరళ పదాలు

**Fig:3 Original image**

The Fig: 4 Minimum and Maximum pixel intensity information for the original image i.e binary image

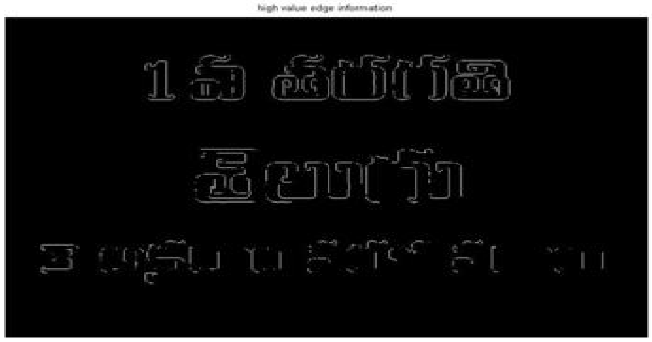


**Fig: 4 The Binaryimages.**

Fig 5 represents the extracted particular word from the original image i.e THARAGATHI

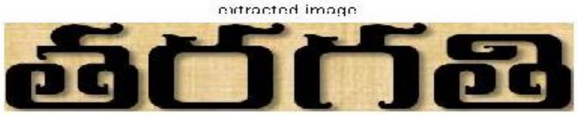


**Fig: 5 Extracted word**

The fig 6 and Fig 7represents the extracted word image and database image here the numbers of matches are reduced due to dissimilar images. In this case, the distance ratio is 0.5

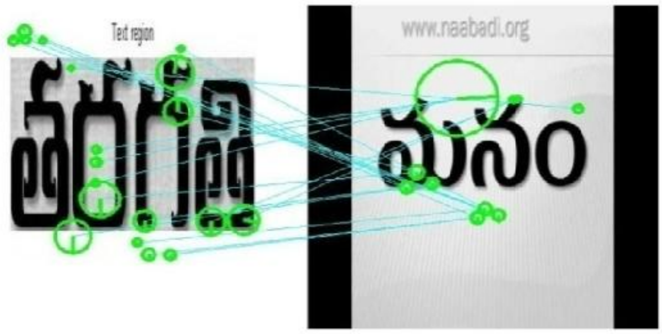


**Fig:6** Kernel features matches for two different images

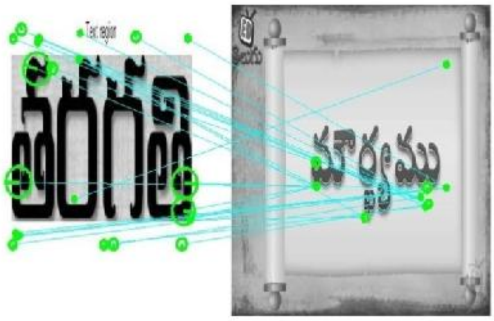


**Fig:7** Kernel features to matching, for dissimilar images

Fig 8 represents the query word image and database image due to similar images the maximum number of matches are found in this case the distance ratio is 0.9

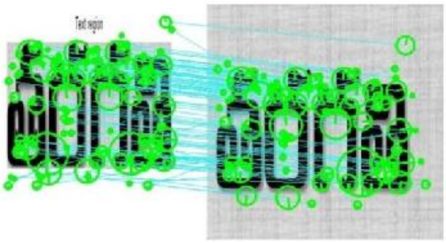


**Fig:8 KERNEL features tomatch,query image, and database image**

## V. CONCLUSION

In this work, we have implemented an efficient method to match aparticular word from Telugu text images. The exploratory outcome demonstrates the vigor of proposed strategy about KERNEL elements. this algorithm uses kernel descriptors and AKD tree tomatch the descriptors. The experimental results show the robustness of the algorithm and less difficult than SIFT,SURF highlights. We need to proceed with this examination work to improve the matching ratio for different scales and revolution images by adopting new methods and also want to differentiate with statistical features and compare with state-of-art methods on large database images.

## References

[1] B. Verma, M. Blumenstein, S. Kulkarni, Recent achievements in off-line handwriting recognition systems, School of Information Technology, Griffith University, Gold Coast Campus.

[2] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, A Bilingual OCR for Hindi-Telugu Documents and its Applications, Centre for Visual Information Technology, International Institute of Information Technology, Hyderabad.

[3] N. Shobha Rani, T. Vasudev, “A Generic Line Elimination Methodology using Circular Masks for Printed and Handwritten Document Images “, Emerging research in computing, information, communication and applications ELSEVIER science and technology, Vol. 3(1), 2014, pp. 589-594.

[4] Rinki Singh, Manideep Kaur, OCR for Telugu Script Using Back-Propagation Based Classifier, International Journal of Information Technology and Knowledge Management, Vol. 2, No. 2, 2010, pp. 639-643.

[5] Suman V Patgar, Vasudev T, Murali S, A system for detection of fabrication in photocopy document, Journal of Computer Science & Information Technology, Vol. 5, No. 14, 2015, pp. 29–35.

[6] N. Sharma, S. Chanda, U. Pal and M. Blumenstein, Word-wise Script Identification from Video Frames, In Proc. ICDAR, 2013

[7] N.sharma, U.Pal, M. Blumenstein, A Study on Word Level Multi-script Identification from Video Frames, Proc. lJCNN, 2014.

[8] P. Shivakumara, N. Sharma, U. Pal, M. Blumenstein, and C. L. Tan, Gradient-Angular-Features for Word-wise Video Script Identification, In Proc. ICPR, 2014

[9] Danish Nadeem, Saleha Rizvi, “Character recognition using template matching”, Project report, Department of Computer Science, Jamia Millia Islamia, New Delhi, 2015.

[10] Nikhil Rajiv Pai, Vijaykumar S. Kolkure, Design and implementation of optical character recognition using template matching for multi fonts /size, International Journal of Research in Engineering and Technology, Vol. 4, No. 2, 2015, pp. 398-400.

[11] Mo Wenying, Ding Zuchun, A Digital Character Recognition Algorithm Based on the Template Weighted Match Degree, International Journal of Smart Home, Vol.7, No. 3, 2013, pp. 53-60.

[12] Jatin M Patil, Ashok P. Mane, Multi Font And Size Optical Character Recognition Using Template Matching, International Journal of Emerging Technology and Advanced Engineering, Vol. 3, No. 1, 2013, pp. 504-506.

[13] Soumendu Das, Sreeparna Banerjee, An Algorithm for Japanese Character Recognition, International Journal of Image, Graphics and Signal Processing (IJIGSP), Vol. 7, No. 1, 2014, pp. 9-15.